HEART FAILURE
Prediction Model

TERM PAPER

Submitted to

Mr. Rishi Rajan Sahay, Assistant Professor,

Shaheed Sukhdev College of Business Studies

By

Kritika Sharma

DECLARATION

I, Kritika Sharma, declare that this project titled “heart failure” is the original work done by me under the guidance of Dr. Rishi Rajan Sahay Professor, Shaheed Sukhdev College of Business Studies, University of Delhi. I further declare that it is made by me as a part of my Certificate course in Data Analytics & Business Intelligence.

Date: 10/01/2024

Name of the student: Kritika Sharma

ACKNOWLEDGEMENT

I would like to express my special thanks of gratitude to my teacher Dr. Rishi Rajan Sahay who gave me the golden opportunity to do this wonderful project. This opportunity helped me in doing a lot of research and discovering many new things.

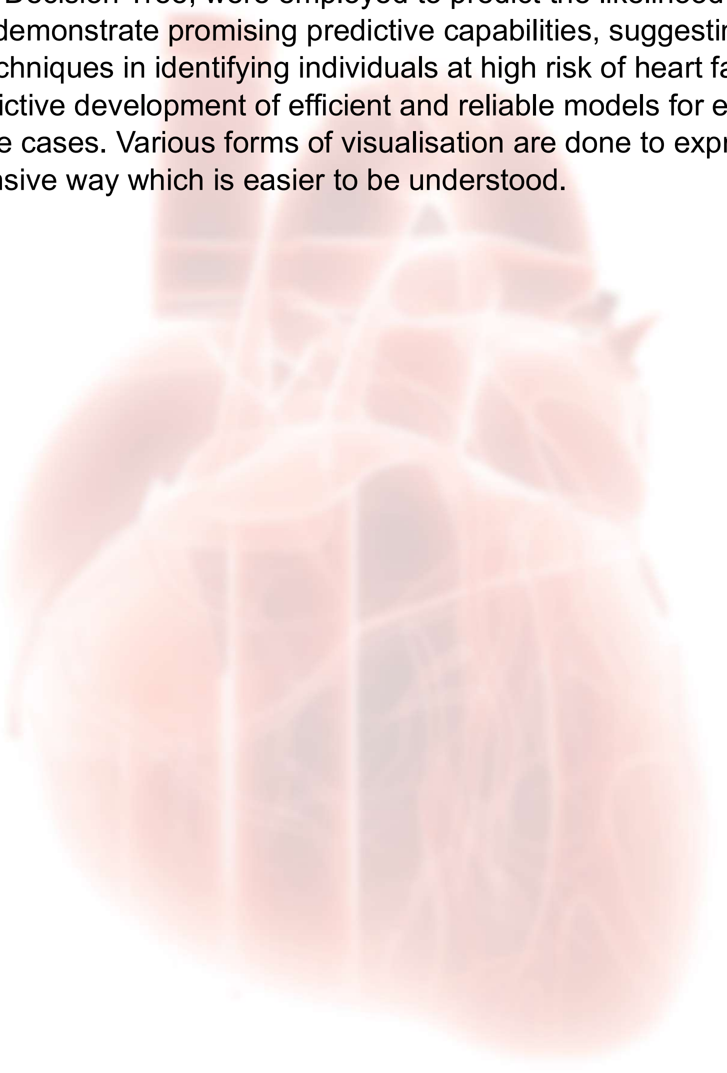
I am overwhelmed in all humbleness and gratefulness to acknowledge my depth to all those who have helped me to put these ideas, well above the level of simplicity and into something concrete.

Thanking you,

Kritika Sharma

ABSTRACT

Heart failure is a critical medical condition associated with a significant mortality rate worldwide. However, with the pandemic i.e. Covid-19 outbreak in 2020, which adversely effected all the spheres of life and economy- the Heart attacks have become the most common and major contributor to the increased death rates in the world. Hence, early detection and accurate prediction of heart failure risk factors play a pivotal role in effective patient care and management. This study explores the application of machine learning algorithms for predicting heart failure based on various clinical features and patient data. A comprehensive dataset containing anonymized patient information, including age, sex, medical history, various biomarkers, and clinical examination results, was utilized for model training and evaluation. Several machine learning algorithms, including Logistic Regression, K-Nearest neighbour, Random Forest, Naïve Bayes and Decision Tree, were employed to predict the likelihood of heart failure occurrence. The outcomes demonstrate promising predictive capabilities, suggesting the potential of machine learning techniques in identifying individuals at high risk of heart failure. This research contributes to the predictive development of efficient and reliable models for early detection and intervention in heart failure cases. Various forms of visualisation are done to express the data in a more comprehensive way which is easier to be understood.



INTRODUCTION

Heart failure also known as congestive heart failure cand be defined as:

> **“***a syndrome, a group of signs and symptoms, caused by an impairment of the heart's blood pumping function.***”**

It is a critical medical condition characterized by the heart's inability to pump blood efficiently, leading to inadequate oxygen supply to various organs and tissues.

Heart failure has always been one of the major causes of death in humans. But after the Corona-virus outbreak, its prevalence continues to rise globally, posing significant challenges to the well-beings of people and the healthcare systems. It is very crucial to identify and predict the risk factors of heart failure to ensure effective & efficient preventive measures, right policies, healthcare facilities and reduce the death rate by the same.

In the context of the modern healthcare, development in science & technology gave birth to data analytics that has emerged as a powerful tool which played a pivotal role in healthcare by revolutionizing disease prediction, risk assessment and management. Leveraging advanced techniques of data analysis in computation of large-scale datasets helps in delving into extensive information of the people which in turn helps in predicting the likelihood of heart failure occurrence.

The aim of this project is to analyse the factors which contribute to the increased risk of the heart failure with the help of the data visualization tools to provide better understanding of the data. The goal is also to establish the relationship between various variables both categorical and continuous to study the impacts of the same on the life expectancy of the individuals. The machine learning algorithms are utilised coupled with the dataset available online to build and evaluate the models that can help in assisting the professionals in personalized healthcare, early interventions and targeted preventive regulations and strategies, thereby enhancing patient care outcomes resulting in reduction in burden of heart failure on healthcare systems.

REASEARCH OBJECTIVES

The primary objective is to identify as well as analyse the most influential features associated with heart failure including demographic information, medical history, biomarkers, and diagnostic measurements and to provide a tool that aids in early detection and prognosis of heart failure. Implement and evaluate various machine learning algorithms, such as Logistic Regression, Random Forest, Gradient Boosting, etc. to develop insights and formulate early intervention strategies & patient care for prevention of heart attacks effectively and efficiently by analysing the predictive features analysed by the models. Compare the performance of different machine learning algorithms to identify the most effective model in predicting heart failure risk.

RESEARCH PROBLEM

Following are the focusses of the research problem:

- To investigate the correlation between the various dependent and independent variables and their impact
- Identifying and selecting the most relevant features or risk factors for heart failure prediction from a pool of potential variables
- Developing a model with high accuracy to predict the likelihood of heart failure based on medical data.
- Ensuring the model's generalizability across diverse patient populations and healthcare settings.

RESEARCH SCOPE & METHODOLOGY

- Obtain a comprehensive dataset comprising patient records, including demographics, medical history, vital signs, and diagnostic tests.
- Cleanse and preprocess the data by handling missing values, scaling, and encoding categorical variables.
- Select relevant features and engineer new ones based on medical domain knowledge.
- Build and evaluate multiple machine learning models such as logistic regression, random forests, etc for heart failure prediction.
- Ensure compliance with data protection laws and ethical guidelines while handling patient data.

POPULATION OF STUDY

The study involve history of patients diagnosed with conditions, covering various demographics, risk factors, and medical histories from multiple healthcare centers or hospitals.

LITERATURE REVIEW

The paper titled “Heart Disease Prediction using Exploratory Data Analysis” by R. Indrakumar & Soumya Ranjan Jena (Assistant Professors) and T.Poongodi (Associate Professor) carried out analysis using a publicly available data for heart disease containing attributes such as age, chest pain type, blood pressure, blood glucose level, ECG in rest, heart rate and four types of chest pain.

Analysis was done with the help of visualization tool like Tableau and K-means clustering.

K-means clustering is an unsupervised class of machine learning algorithm. Usually, unsupervised algorithms project the desired output without referring any value. In K-means clustering algorithm. The goal of clustering is to divide the population or set of data points into a number of groups so that the data points within each group are more comparable to one another and different from the data points within the other groups. It is essentially a grouping of things based on how similar and different they are to one another.

Here, K-means clustering algorithm was selected because of its ‘efficiency, simplicity, capacity to produce even sized population and scalability in handling the web dataset to produce accurate output.’ K-means algorithm have minimum sum of squares to categorize clusters of data points. Here the dataset had 209 observations of 7 variables. The initial center of cluster is computed with the following steps:

- Identify random K clusters
- Iteratively find the significant clusters
- If the distance between the observation and its nearest cluster center is higher than the distance among other closest cluster centers then the observation is replaced with nearest centre by calculating Euclidean distance among the cluster and the observation.
- Within cluster sum of squares is calculated as:

$$\sum_{k=1}^{K} \sum_{i \in S_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$$

where S_k is the set of observations in the kth cluster and $\bar{x}_{kj}$ is the jth variable of the cluster center for the kth cluster.

The iteration will get stop if the difference between the sum of squares in two successive iterations is minimal and this is called Final Cluster Centers.

The variables considered to predict the heart disease are age, maximum heart rate, chest pain type and disease. Here, four types of chest pains are considered.

Tableau uses centroid-based k-means clustering algorithm that divides the data into K-number of clusters. Dashboards are created with the data set after applying K-means algorithm. It provides visual appealing clusters in order to predict the occurrence of heart disease from the given dataset.

It was found that the target class with the age ranging from 50 to 55 have high risk of heart disease as the development of coronary fatty streaks starts in this age range. It was also deduced that population with diabetes and high blood pressure is expected to get heart disease as compared to the ones who doesn’t have any.

It was concluded that heart stroke and vascular disease are the major cause of disability and premature death and chest pain is the key to recognize the heart disease. In the following paper, the heart diseases are predicted by considering major factors with four types of chest pain.

STUDY OF CORRELATION

Correlation can be defined as:

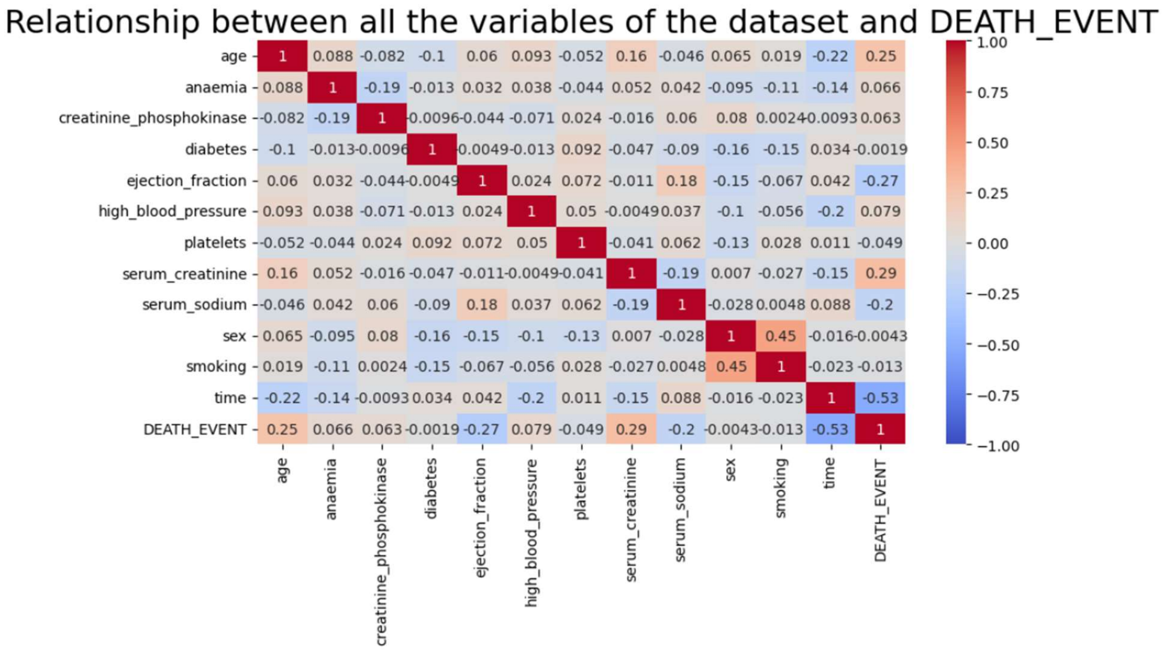
“*The statistical measure that indicates the extent to which two or more variables change together*.”

It helps in understanding the relationship between variables and how they might influence each other. Correlation can be positive, negative, or zero.

- **Positive correlation**: When one variable increases, the other variable tends to increase as well. A correlation coefficient close to +1 indicates a strong positive correlation.
- **Negative correlation**: When one variable increases, the other variable tends to decrease. A correlation coefficient close to -1 indicates a strong negative correlation.
- **Zero correlation**: There's no apparent linear relationship between the variables. A correlation coefficient close to 0 suggests no linear relationship between the variables.

Correlation helps in various ways:

1) **Identifying Relationships**: It assists in understanding how changes in one variable might be associated with changes in another. For instance, identifying if high blood pressure correlates with an increased likelihood of heart failure.
2) **Feature Selection**: It aids in selecting relevant features for predictive modelling. Highly correlated features might provide redundant information, so understanding these relationships helps in feature selection for models.
3) **Assumptions in Analysis**: Correlation analysis is essential for various statistical analyses and machine learning algorithms. For instance, in linear regression, correlated predictors might violate the assumption of multicollinearity.



Heatmap is useful as it helps us to show relationship using colours and annotations. It also gives us an idea of which variables to choose to train the model or algorithms to get better results.

<u>Following observations are made:</u>

We notice that there is-

- positive correlation between DEATH_EVENT and serum creatinine i.e.0.29 and age i.e.0.25.
- negative correlation between DEATH_EVENT and time, ejection fraction and serum sodium

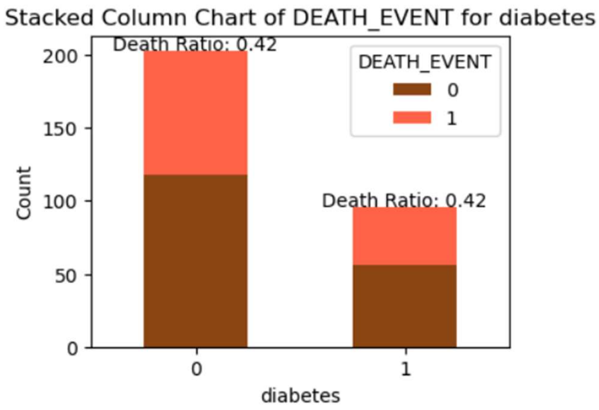
  i.e. -0.53, -0.27 & -0.2 respectively.

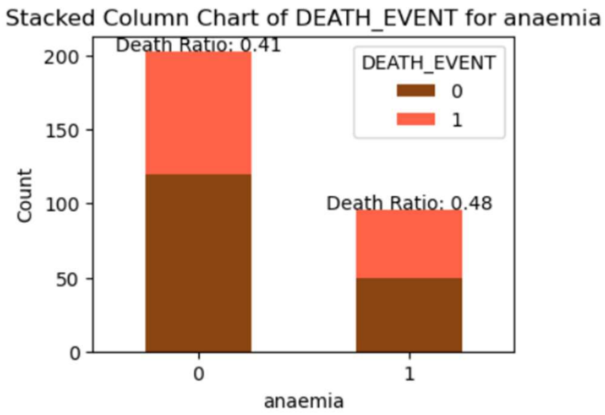
DATA ANALYSIS

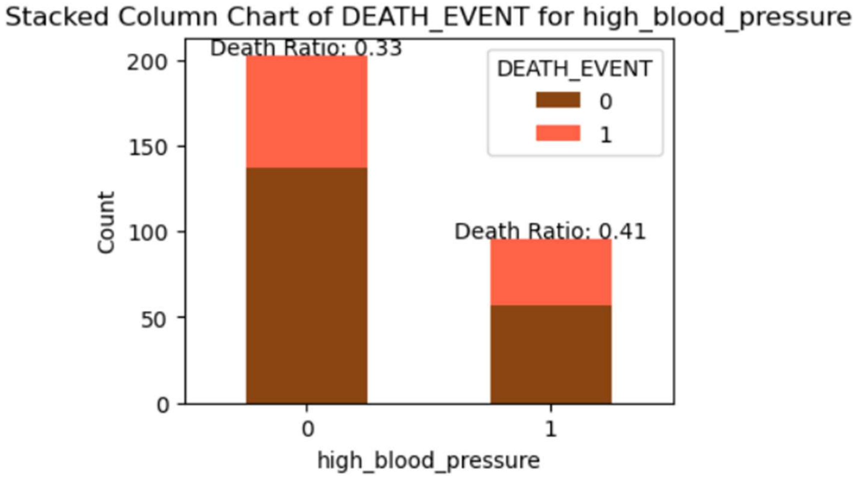
DATA VISUALIZATION

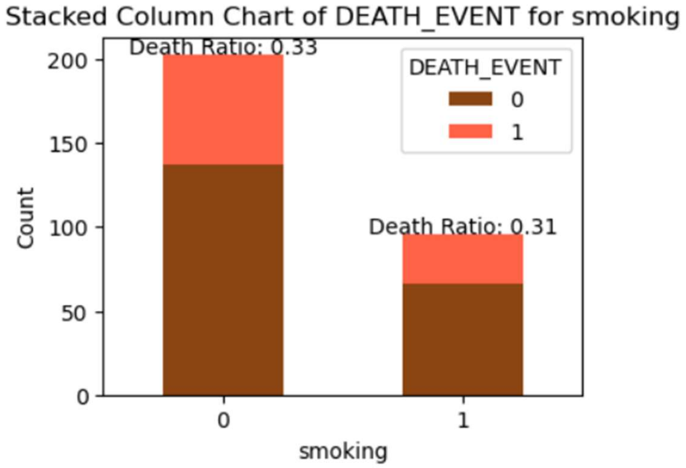
Data visualization is a powerful tool that enhances data exploration, aids in the discovery of patterns and insights, and effectively communicates complex information, thereby playing a pivotal role in the data analysis process.

- <u>Impact of categorical variables on Death Event</u>





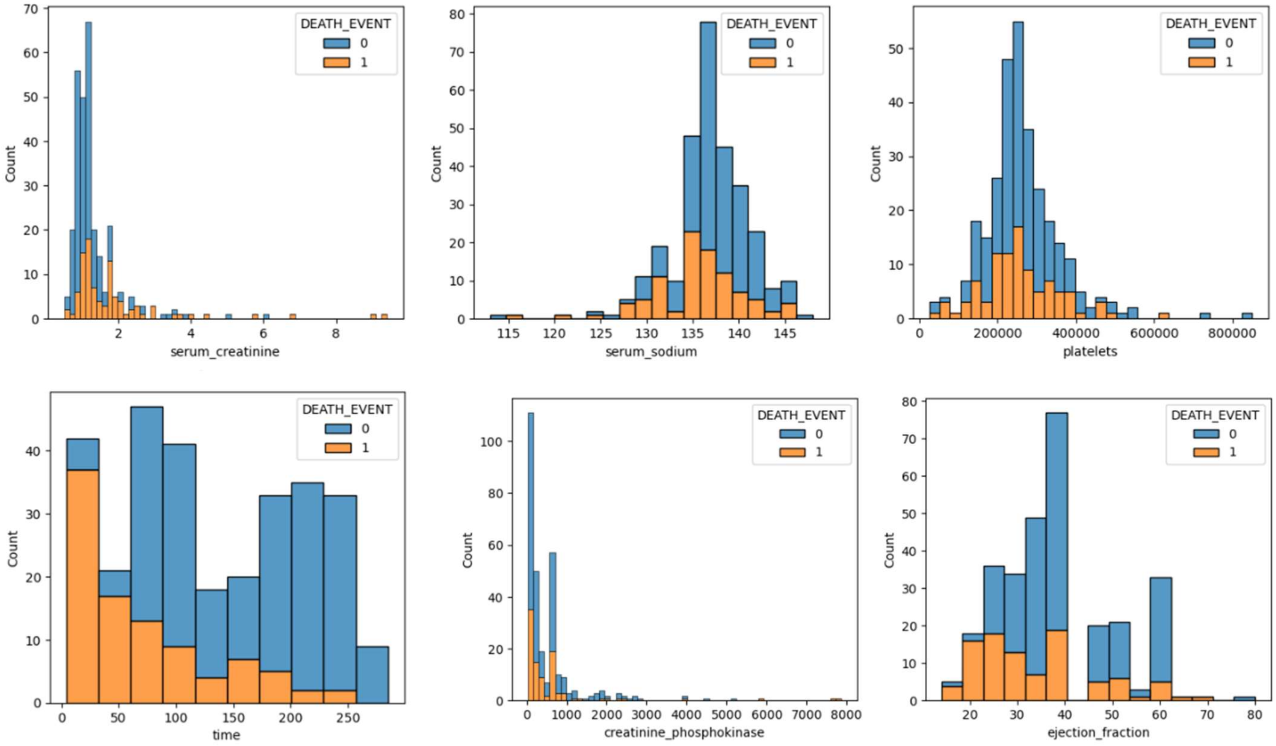




Above graphs tell the relationship between various categorical variables & the death event when death event is 0 or 1 respectively.

It can be observed that anaemia, high blood pressure, diabetes are all the factors which adversely impact the health of the individuals.

- <u>Impact of continuous variables on Death Event</u>



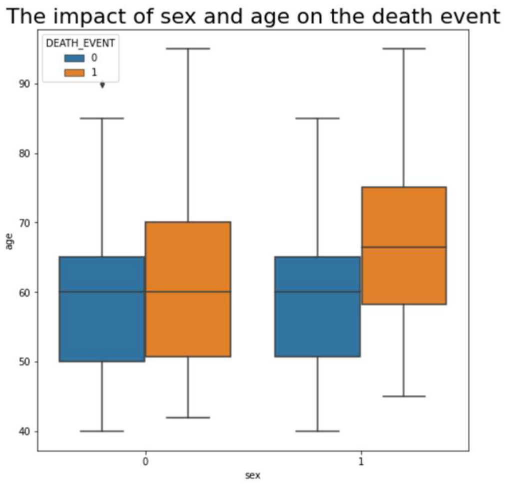
The increased concentration of creatinine phosphokinase & serum sodium are dangerous for the human body which may disbalance & effect the proper functioning of the heart and heart vessels.

- <u>impact of gender and age on the death event</u>

Boxplots serve as an essential exploratory data analysis tool, providing a concise summary of the distribution of numerical data while allowing for easy comparison and identification of potential anomalies or patterns within datasets.

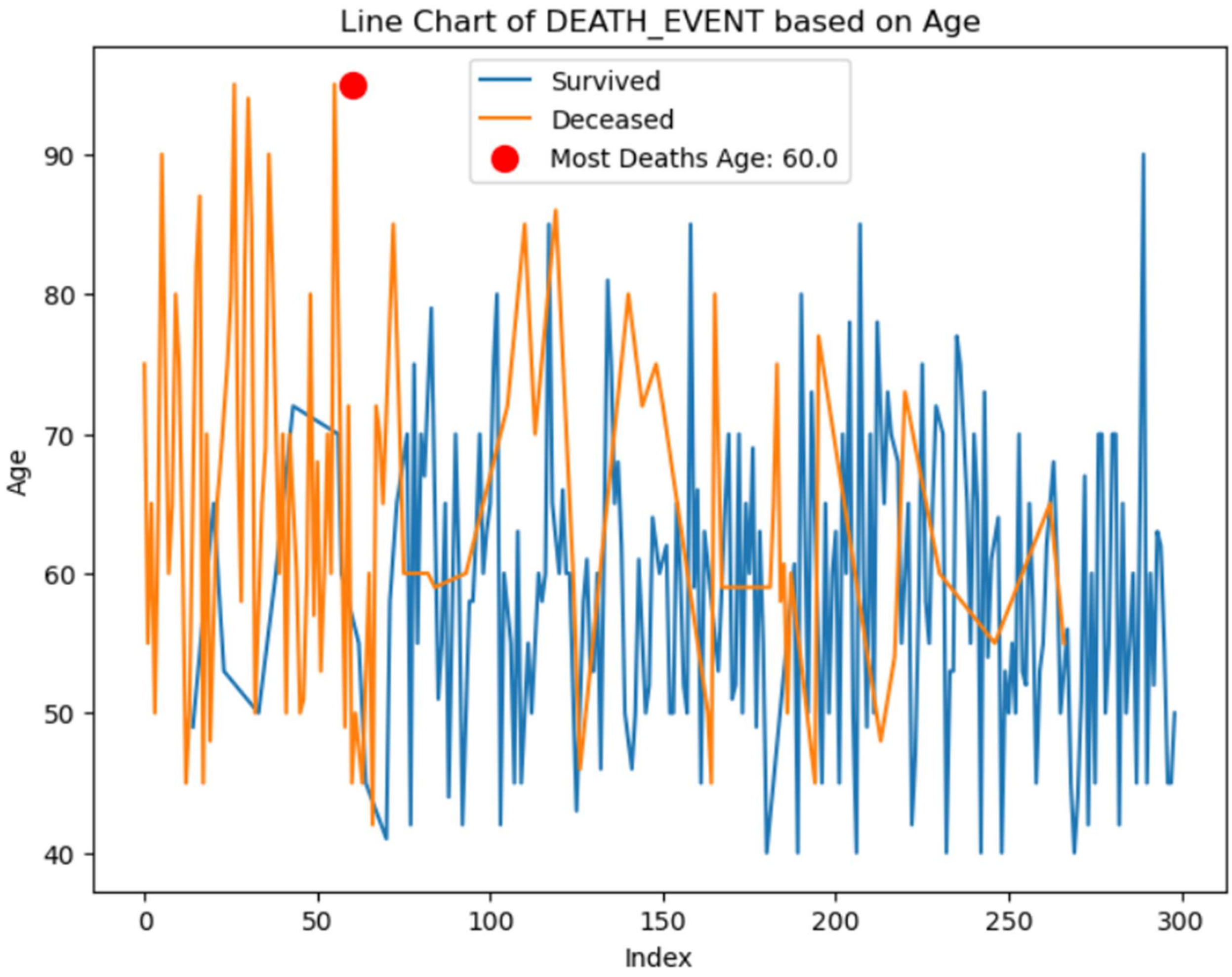
Boxplots depict the median, quartiles (25th and 75th percentiles), and overall range of the dataset, providing a visual summary of the data's central tendency and dispersion. They highlight potential outliers by showing data points beyond the whiskers, enabling the identification of extreme values that might significantly affect statistical analysis.

Boxplots facilitate the comparison of distributions between multiple groups or categories, helping analysts observe differences in central tendency and variability among various subsets of data. They display data symmetry, skewness, or asymmetry. The positioning and length of the box and whiskers indicate the skewness of the distribution.

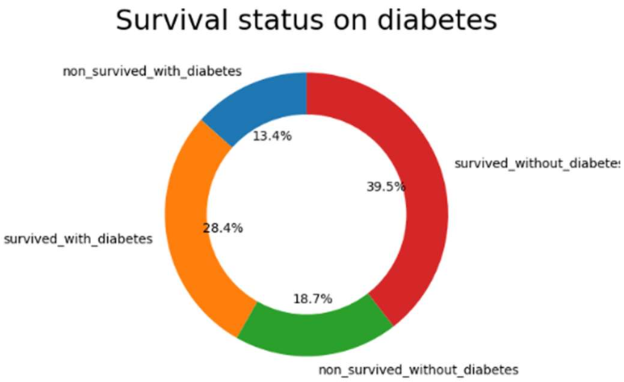


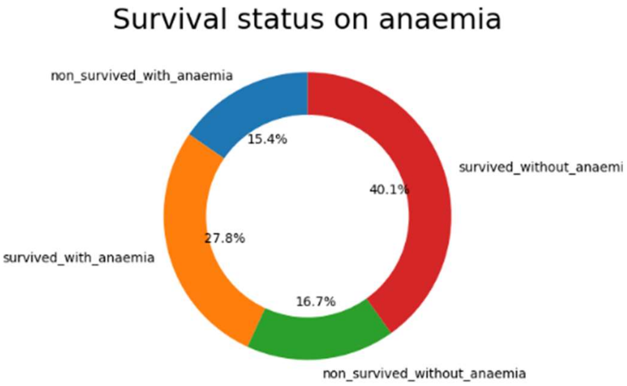
It can be observed that for males (sex=1) the median is 65 when death event= 1 & about 60 when death event =0. However, when it comes to females the median is 60 and doesn't change whether the death event is 0 or 1.

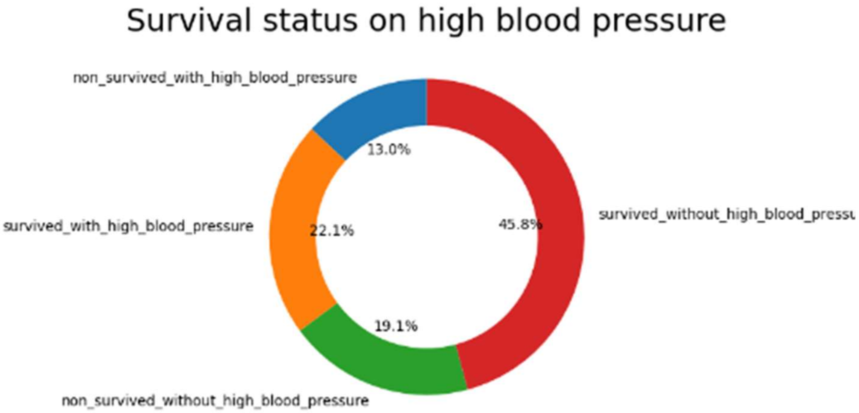
Same can be concluded from the line chart as well.



SURVIVAL STATUSES







Above doughnut charts provide a glimpse of the survival status of the individuals in cases where they have a certain condition like high blood pressure and anaemia or when they are active smokers or not.

It can be fairly deduced that individuals with healthy habits and body have fair chances of living a long and healthy life with lesser risk to the heart attacks as compared to people who are active smokers and have health issues.

DATA MODELLING & PREDICTION

Various classifiers are taken into consideration like logistic regression, K-nearest neighbors, decision tree, naïve bayes & random forest.

Following Sklearn packages were installed:

1) Train-test split
2) Standard scaler
3) Logistic regression
4) K-Neighbors classifier
5) Decision tree clssifier
6) Gaussian NB
7) Random forest classifier
8) Accuracy score

Logistic Regression

Logistic regression is a supervised machine learning algorithm mainly used for binary classification where we use a logistic function, also known as a sigmoid function that takes input as independent variables and produces a probability value between 0 and 1.

K-Nearest Neighbors

K-Nearest Neighbor (KNN) is a supervised machine learning algorithm based on the principle of similarity.

Random Forest

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in Machine Learning.The algorithm is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model

Decision Tree

It is a versatile supervised machine-learning algorithm, which is used for both classification and regression problems. It is one of the very powerful algorithms. And it is also used in Random Forest to train on different subsets of training data, which makes random forest one of the most powerful algorithms in machine learning.

Python Code used:

- Decision Tree Classifier Initialization:

```
dt_model = DecisionTreeClassifier(criterion="entropy", max_depth=2)
```

Initializes a Decision Tree Classifier with specific parameters:

criterion="entropy": Uses the entropy criterion to measure the quality of a split.

max_depth=2: Specifies the maximum depth of the tree to be 2, limiting the depth to control overfitting.

- Model Training:

```
dt_model.fit(x_train_scaled, y_train)
```

Fits the Decision Tree Classifier model using the scaled training data (x_train_scaled) and corresponding target values (y_train).

- Making Predictions:

```
dt_prediction = dt_model.predict(x_test_scaled)
```

Uses the trained model to make predictions on the scaled testing data (x_test_scaled).

- Calculating Accuracy:

```
dt_accuracy = (round(accuracy_score(dt_prediction, y_test), 4) * 100)
```

Computes the accuracy score of the model's predictions by comparing them against the actual target values (y_test).accuracy_score is a function that measures the accuracy of classification algorithms. It compares predicted labels (dt_prediction) with actual labels (y_test) and returns the accuracy score.

The accuracy score is multiplied by 100 and rounded to four decimal places (round(accuracy_score(dt_prediction, y_test), 4) * 100).

The resulting accuracy percentage is stored in the dt_accuracy variable.

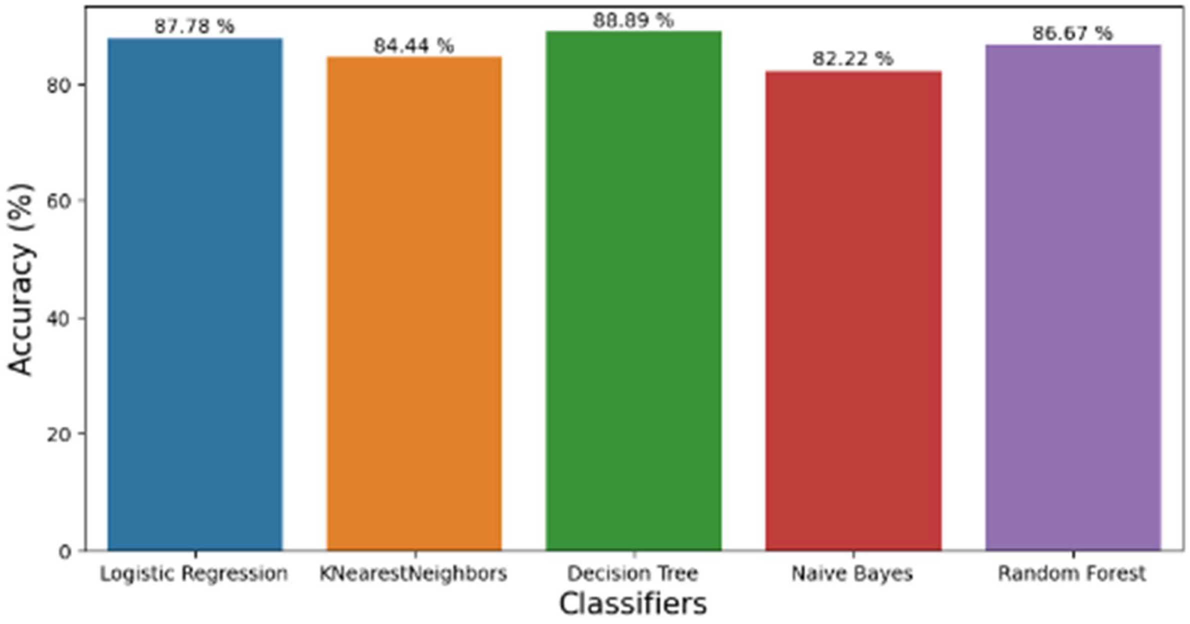
- Appending Accuracy to List:

```
accuracy_list.append(dt_accuracy)
```

Appends the calculated accuracy percentage (dt_accuracy) to an existing list named accuracy_list, which likely stores accuracy scores from multiple models or iterations for further analysis or comparison.

Now lets look at the accuracy score of different models:

Accuracy of different classifiers



Therefore, The model with the least score is Naïve Bayes with 82.22%.

Therefore Naïve Bayes cannot be used.

The model here isnt binary or lineary separable so logistic regression cannot be used in such case.

Here, data isnt clustered therfore using KNN is not a good choice as well.

Hence the only possible option is Decision tree or Random Forest.

It can be seen that Decision tree is the model with the most precision & accuracy i.e. 88.9%. Random forest has an accuracy score of 85.56%.

So the Decision tree model is the perfect choice in the prediction process of heart failure data with highest accuracy & robustness.

CONCLUSION

This project regarding data analysis focussed on prediction of heart failure based on the dataset available online. Various features and relationship among different variables were examined through EDA, modelling and predictive analytics and insights were developed. The comparison between different models were done theoretically and the best model was decided for this project based on the usage as well as the accuracy score of the models. The model with the highest accuracy score was taken into account.

This project provided valuable insights into the most influential factors contributing to heart failure occurrence, such as age, serum creatinine levels, ejection fraction, and comorbidities like diabetes and hypertension.

Factors like diabetes, anaemia, High BP showed a more prominent influence on heart failure occurrence. On the contrary, gender seemed to be less impactful.

The findings of the study emphasized the need of monitoring of critical factors such as ejection fraction & serum creatinine levels in identifying higher risk of heart failure in individuals. As well as focus on early detection and effective healthcare strategies should be given great attention.

References

1) "Heart Disease Prediction using Exploratory Data Analysis" by R. Indrakumar & Soumya Ranjan Jena (Assistant Professors) and T.Poongodi (Associate Professor)

2) Various online platforms e.g.. Google, Kaggle etc

3) Class notes